MULTIVIEW PEDESTRIAN LOCALISATION VIA A PRIME CANDIDATE CHART BASED ON OCCUPANCY LIKELIHOODS

Yuyao Yan[†] *Ming Xu*[⋆] *Jeremy S. Smith*[†]

[†] Department of Electrical Engineering and Electronics, University of Liverpool, L69 3BX, Liverpool, UK
[⋆] Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China

ABSTRACT

A sound way to localize occluded people is to project the foregrounds from multiple camera views to a reference view by homographies and find the foreground intersections. However, this may give rise to phantoms due to foreground intersections from different people. In this paper, each intersection region is warped back to the original camera view and is associated with a candidate box of the average pedestrians' size at that location. Then a joint occupancy likelihood is calculated for each intersection region. In the second step, essential candidate boxes are identified first, each of which covers at least a part of the foreground that is not covered by another candidate box. The non-essential candidate boxes are selected to cover the remaining foregrounds in the order of the joint occupancy likelihoods. Experiments on benchmark video datasets have demonstrated the good performance of our algorithm in comparison with other state-of-the-art methods.

Index Terms— image motion analysis, object detection, image fusion, visual surveillance

1. INTRODUCTION

An important task in video surveillance is to detect multiple pedestrians. These pedestrians may be partially occluded by each other in a camera view. To overcome this problem, multiple cameras can be deployed to provide complementary information about the moving targets, because the overlapped pedestrians in one camera view may be separated in another camera view. The information provided by the multiple cameras is able to make detection more robust and accurate. When working with multiple camera views, homography has been widely used for the association and information fusion of multi-camera observations. Khan and Shah [1] projected the foreground likelihoods from individual camera views to a reference view by using ground-plane homographies and identified the intersection regions as the locations of pedestrians. This method avoids integrating the features extracted in individual camera views, as the latter is vulnerable to the occlusion of pedestrians. This approach adds robustness to the detection of pedestrians in moderate density. However, the foreground projection of one pedestrian from a camera view may falsely intersect with that of a non-corresponding pedestrian from another camera view, which leads to phantoms in pedestrian detection.

Significant research has been undertaken to avoid the generations of phantoms. Khan and Shah [2] extended their early work by projecting the foreground likelihoods from individual camera views to a reference view by using the homographies of a set of parallel planes and selected the most heavily overlapping areas as the pedestrian locations. Eshel and Moses [3] positioned the cameras at high locations so that they were looking downwards, which can reduce occlusions. In addition, the intensities projected from multiple camera views to the same location of a reference view are correlated, which can reduce phantoms.

Multiview pedestrian detection is sometimes thought of as an optimization problem. Ge et al. [4] proposed a generative sampling-based approach that models a pedestrian as an upright cylinder. Gibbs sampling is used to estimate the number and the locations of pedestrians in a crowd. Akos and Benedek [5] extended the classical Bayesian Marked Point Process (MPP) model [6] to a 3DMMP model which utilizes the pixel-level features from pedestrians' heads and feet, instead of the whole silhouettes, to reduce the number of phantoms. Fleuret et al. [7] calculated a probabilistic occupancy map (POM) in the ground plane which is divided into grids. A pedestrian is modeled as a rectangle of the average size of pedestrians standing in each grid. Then an iterative algorithm is utilized to find the optimal rectangles which cover more foreground pixels and less background pixels in both camera views. Peng et al. [8] modeled each pedestrian as a rectangle similar to [7] and analyzed the occlusion relationship among such rectangles to identify phantoms by using a Bayesian network model.

In this paper an algorithm is proposed for multiview pedestrian localisation. The foregrounds from two camera views are warped to a top view using homographies. Then each intersection region is warped back to both camera views.

Each warped back region is associated with a candidate box standing on that region and of the average size of pedestrians. The joint occupancy likelihood of each candidate is calculated by taking into account the foreground likelihood and the observability of the candidate boxes in both camera views. At the second stage, a prime candidate chart is developed to select the essential candidates, each of which covers at least a foreground region that is not covered by another candidate. Afterwards the non-essential prime candidates are selected to cover the remaining foreground regions in terms of the joint occupancy likelihoods.

The contributions of this paper are twofold: the use of the prime candidate chart greatly reduces the search space of the optimized solution; the joint occupancy likelihood considers the foreground likelihood and the observability of each candidate.

2. FOREGROUND SEGMENTATION

Background subtraction is used for the foreground detection in each camera view, in which the colour of each pixel is modelled as a Gaussian mixture model [9]. In the real world, people may be, or appear to be, walking side by side. This complicates the foreground projections from multiple camera views. In this paper the convex hull of each foreground region is used to separate such pedestrians. The spaces between the contour and the convex hull are defined as convexity defects. Each convexity defect has three main points: the start point p_s, the end point p_e and the farthest defect point p_d.

The convex hull of a group of side-by-side pedestrians usually have one or more large convexity defects facing upwards and between their heads. In order to locate the convexity defects, the direction of each convexity defect is calculated as the bisector of the angle $\angle p_s p_d p_e$:

$$\beta = \arctan\left(\frac{\overrightarrow{p_d p_s}}{|\overrightarrow{p_d p_s}|} + \frac{\overrightarrow{p_d p_e}}{|\overrightarrow{p_d p_e}|}\right) . \quad (1)$$

By thresholding the area of the convexity defect triangles and limiting the angle of β from $-\frac{\pi}{6}$ to $\frac{\pi}{6}$, which ensures the convexity defect is facing upwards, the farthest defect points can be identified and the side-by-side pedestrians are split at that location. The same process is recursively applied to the split foreground regions so that more than two side-by-side pedestrians in a group can be separated.

3. HOMOGRAPHY ESTIMATION

Planar homography is defined by a 3×3 transformation matrix between a pair of captured images of the same plane with two cameras. Let $\mathbf{x}^c$ and $\mathbf{x}^t$ be the homogeneous coordinates of a point in camera view c and its corresponding point in a virtual top view. They are associated by the homography matrix $\mathbf{H}^{t,c}$ as $\mathbf{x}^c \cong \mathbf{H}^{t,c}\mathbf{x}^t$. After each camera is calibrated, a 3×4 projection matrix can be calculated using the intrinsic and extrinsic parameters of the camera: $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \mathbf{m}_4]$. Then the homography matrix for the ground plane is:

$$\mathbf{H}_0^{t,c} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_4] . \quad (2)$$

The homography matrix for a plane parallel to and at a height of h above the ground plane is:

$$\mathbf{H}_h^{t,c} = \mathbf{H}_0^{t,c} + [\mathbf{0}|h\mathbf{m}_3] , \quad (3)$$

where $[\mathbf{0}]$ is a 3×2 zero matrix [10].

In this paper the foregrounds in the individual camera views are projected to the top view using the homographies for the waist plane. Then the foreground intersections are warped back to the individual camera views by using the ground-plane homographies. The warped region for a pedestrian is ideally located at the bottom of the foreground region. If it is well below the bottom of the foreground region, it is a phantom; If it is above the bottom, it may be either a phantom or a pedestrian standing behind another.

4. JOINT OCCUPANCY LIKELIHOODS

Suppose there are N cameras. F_i represents the foreground observation in camera view i. For a specific intersection region I in the top view, there are N warped back intersection regions $\{I_1, I_2, \ldots, I_N\}$, each of which is cast in an individual camera view and is associated with a rectangular box A_i of the average size of pedestrians who are standing there. Let X be the event that there is a pedestrian at intersection region I in the top view. Given foreground observations $F_1, F_2, \ldots, F_N$, we are interested in finding the posterior probability of event X happening. Three independent measurements derived from each foreground region are the foreground pixel set f, the foot location d and the height observation h.

By using Bayes law and considering the conditional independence between the three measurements, we have:

$$\begin{aligned} &P(X|F_1, F_2, \ldots, F_N) \propto P(F_1, F_2, \ldots, F_N|X)P(X) \\ &\propto \prod_{i=1}^{N} P(F_i|X) = \prod_{i=1}^{N} P(f_i, d_i, h_i|X) \\ &= \prod_{i=1}^{N} [P(f_i|X)P(d_i|X)P(h_i|X)] . \end{aligned} \quad (4)$$

f_i is the foreground pixel set enclosed by candidate box A_i, i.e. $f_i = F_i \cap A_i$. $P(f_i|X)$ can be approximated by the foreground pixel ratio:

$$P(f_i|X) = \frac{number\ of\ foreground\ pixels\ in\ A_i}{number\ of all\ pixels\ in\ A_i} . \quad (5)$$

d_i is used to measure the distance between the bottom of the candidate box and that of the corresponding foreground box in camera view i. Suppose the vertical coordinate of the

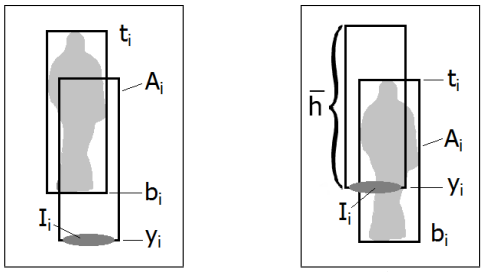
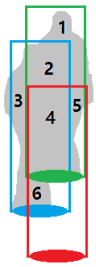


Fig. 1. A schematic diagram of the variables related to the estimation of d_i and h_i.



(a)

Sub-region:	1	2	3	4	5	6
Red	+	+	+	X	X	X
Green	X	X	+	X	X	+
Blue	+	X	X	X	+	X

(b)

Fig. 2. (a) Decomposition of a foreground region into sub-regions. (b) The corresponding prime candidate chart.

warped region centroid is y_i with a variance $\sigma^2_{y,i}$ which is determined by the height of the warped region, the vertical coordinate of the foreground region bottom is b_i with a variance $\sigma^2_{b,i}$, and the vertical coordinate of the foreground region top is t_i (see Fig. 1(a)). Then the Mahalanobis distance is:

$$d_i = \begin{cases} 0 & if\ b_i \le y_i \le t_i \\ \frac{(y_i - b_i)^2}{(\sigma^2_{y,i} + \sigma^2_{b,i})} & otherwise \end{cases} . \quad (6)$$

d_i is chi-squared distributed with $n = 1$ degree of freedom, i.e. $d_i \sim \chi^2_1$. Suppose the tail probability on the chi-square distribution is denoted by $Q_{\chi^2}(x, 1) = \int_x^\infty p_{\chi^2}(t, 1)dt$. Given the value of d_i, $P(d_i|X)$ is determined as:

$$P(d_i|X) = Q_{\chi^2}(d_i, 1) . \quad (7)$$

h_i is the maximum height of the pedestrian candidate. It is the distance between the bottom of the candidate box and the top of the corresponding foreground box in camera view i (see Fig. 1(b)). Suppose the heights of adults are Gaussian distributed as $h \sim G(\overline{h}, \sigma^2_h)$ and the tail probability on the Gaussian distribution is denoted by $Q_G(X) = \int_x^\infty p_G(t)dt$. Then the maximum height h_i and $P(h_i|X)$ are defined as:

$$h_i = t_i - y_i \quad (8)$$

$$P(h_i|X) = 1 - Q_G(h_i) . \quad (9)$$

Both d_i and h_i are normalized by the average height $\overline{h}$ of the pedestrians standing at the warped back region.

5. PRIME CANDIDATE CHARTS

The joint occupancy likelihood is derived separately for each pedestrian candidate. To encode the interactivity such as occlusion and grouping between pedestrians, global optimization is carried out for the multiview pedestrian localization. We borrowed the idea from the Quine-McCluskey method [11] for the minimisation of Boolean functions.

Each foreground region is decomposed into sub-regions according to the overlapping relationship of all the candidate boxes associated with that foreground region. Each sub-region must be made as large as possible while ensuring that there is no transition on the overlapping candidate boxes inside the sub-region (see Fig. 2(a)).

A prime candidate chart is introduced to select a minimum set of pedestrian candidates to cover all the foreground sub-regions of interest. In the prime candidate chart (see Fig. 2(b)), the foreground sub-regions in all the camera views are listed across the top of the chart, and the pedestrian candidates are listed down the left side. If a pedestrian candidate covers a given sub-region, an X is placed at the intersection of the corresponding row and column; otherwise, a plus sign is placed at the intersection.

The prime candidate chart is updated as follows:

(1) All the pedestrian candidates are scanned. If the joint occupancy likelihood of any candidate is too low, it is removed from this chart by replacing the X's in the corresponding row by plus signs.

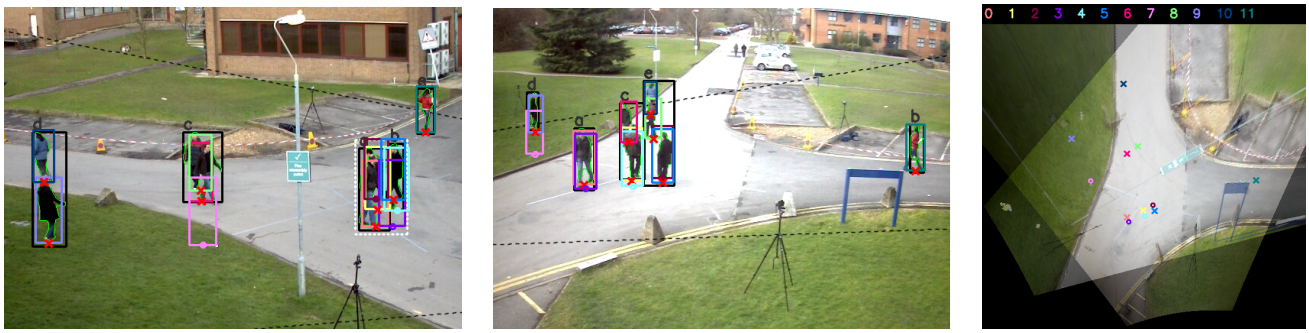
(2) All the sub-regions are scanned. If a sub-region is too small or does not contain a significant portion of foregrounds, it is removed by replacing the X's in the corresponding column by plus signs.

(3) The sub-regions are scanned again. If a foreground sub-region is covered by only one candidate, the candidate is recorded as an essential candidate and recognized as a pedestrian. The X in the corresponding row and column is replaced by a plus sign.

(4) If the essential candidates do not cover all the sub-regions, additional non-essential candidates are selected to cover the remaining sub-regions in the order of their joint occupancy likelihoods. This step is run iteratively until there are no sub-regions left.

6. EXPERIMENTAL RESULTS

To evaluate the proposed algorithm, experiments were performed on the PETS2009 City Center (CC) dataset [12] which is a benchmark dataset containing a crowd of pedestrians in 8 calibrated camera views. Only two camera views (views 1 and 2) were used in our experiments. Each view has 795 frames, in which the first 200 frames were used to train the



(a) (b) (c)

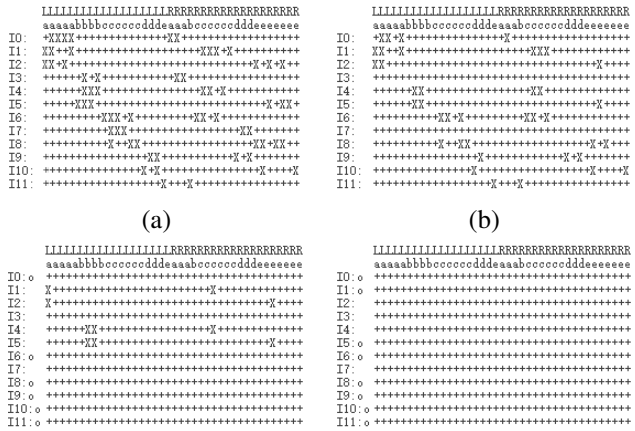
Fig. 3. The detection results at frame 729 on the PETS2009 CC dataset. (a)(b) Camera views 1 and 2, and (c) the top view.

background model and the remaining frames were used to evaluate the performance.

Fig. 3 shows the detection results at frame 729 on the PETS2009 CC dataset. The borderlines of the overlapping field of views are shown as black dashed lines. The contour and bounding box of each foreground region are in green and black, respectively. The original bounding box for split side-by-side pedestrians are in white dotted lines. Each foreground intersection region in the top view, which is represented in a distinguished colour, corresponds to a pair of candidate boxes represented by the same colour in both camera views. The intersection region IDs are shown at the top of Fig. 3(c) and also in the same distinguished colour. An identified pedestrian is labeled with a cross at the bottom of its candidate box, while each phantom is labeled with a circle.

Fig. 4 is the prime candidate chart at the same frame as Fig. 3. Fig. 4(a) shows the initial chart, Fig. 4(b) is the chart after steps 1 and 2 by removing invalid candidates and foreground sub-regions, Fig. 4(c) is that after step 3 by removing essential candidates, and Fig. 4(d) is that after step 4 by selecting non-essential candidates. Down the left side of the chart is the list of pedestrian candidates. If a candidate is identified as a pedestrian, then it is labeled with a circle. At the top of each chart, L and R represent the left and right camera views. In the second row, a-e are foreground region IDs. If a foreground region ID appears successively several times, they refer to the sub-regions decomposed from the same foreground region.

For a performance comparison with some state-of-the-art algorithms, three metrics were evaluated: PRECISION, RECALL and TER (total error rate) [5]. PRECISION and RECALL were defined as the ratios $TP/(TP+FP)$ and $TP/(TP+FN)$, where TP, FP and FN are the numbers of true positives, false positives and false negatives, respectively. For these two ratios, a larger value indicates a better performance. TER is used to measure the detection accuracy, which considers both FP and FN. A lower TER value corresponds to a better performance. The comparison results based on PETS2009 CC dataset are shown in Table. 1. MvBN [8] was evaluated in PRECISION, RECALL and TER. POM [7] and 3DMPP [5] only used TER as the evaluation metric. The



(a) (b) (c) (d)

Fig. 4. The prime candidate chart at frame 729: (a) the initial one, (b) after steps 1 and 2, (c) after step 3, (d) after step 4.

Table 1. Evaluation results on PETS 2009 CC dataset.

Method	RECALL	PRECISION	TER
MvBN	0.90	0.97	0.13
POM	-	-	0.27
3DMPP	-	-	0.31
Proposed	0.97	0.99	0.04

proposed algorithm outperforms these algorithms in terms of PRECISION, RECALL and TER.

7. CONCLUSIONS

We have proposed an algorithm for multiview pedestrian localization, which is based on foreground intersections in a virtual top view. The joint occupancy likelihoods and the prime candidate chart used in this paper add the robustness to the pedestrian localization. Experiment results have shown its better performance than some state-of-the-art algorithms that use 3-4 cameras.

8. REFERENCES

[1] Saad M. Khan and Mubarak Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, 2006, Proceedings, Part IV*, 2006, pp. 133–146.

[2] Saad M. Khan and Mubarak Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 505–519, 2009.

[3] Ran Eshel and Yael Moses, "Tracking in a dense crowd using multiple cameras," *International Journal of Computer Vision*, vol. 88, no. 1, pp. 129–143, 2010.

[4] Weina Ge and Robert T. Collins, "Crowd detection with a multiview sampler," in *Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision*, 2010, pp. 324–337.

[5] Ákos Utasi and Csaba Benedek, "A bayesian approach on people localization in multicamera systems," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 23, no. 1, pp. 105–115, 2013.

[6] Wenjie Ge and Robert T Collins, "Marked point processes for crowd counting," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2913–2920.

[7] François Fleuret, Jérôme Berclaz, Richard Lengagne, and Pascal Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 267–282, 2008.

[8] Peixi Peng, Yonghong Tian, Yaowei Wang, Jia Li, and Tiejun Huang, "Robust multiple cameras pedestrian detection with multi-view bayesian network," *Pattern Recognition*, vol. 48, no. 5, pp. 1760–1772, 2015.

[9] Chris Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, 1999, vol. 2, pp. 246–252.

[10] Jie Ren, Ming Xu, Jeremy S. Smith, and Shi Cheng, "Multi-view and multi-plane data fusion for effective pedestrian detection in intelligent visual surveillance," *Multidimensional Systems and Signal Processing*, vol. 27, no. 4, pp. 1007–1029, 2016.

[11] Willard V Quine, "The problem of simplifying truth functions," *The American Mathematical Monthly*, vol. 59, no. 8, pp. 521–531, 1952.

[12] Anna Ellis, Ali Shahrokni, and James Ferryman, "Pets2009 and winter-pets2009 results: A combined evaluation," in *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*. IEEE, 2009, pp. 1–8.